2000

5. Metacognition and Computer-Based Testing

Gregory Schraw
University of Nebraska - Lincoln

Steven L. Wise
University of Nebraska-Lincoln, wisesl@jmu.edu

Linda L. Roos
University of Nebraska-Lincoln, lroos@unlinfo.unl.edu

Published in ISSUES IN THE MEASUREMENT OF METACOGNITION,
ed. Gregory Schraw & James C. Impara (Lincoln, NE: Buros Institute of
Mental Measurements, 2000).

5

Metacognition and Computer-Based Testing

Gregory Schraw

Steven L. Wise

Linda L. Roos

The University of Nebraska-Lincoln

Metacognition refers to thinking about thinking, or more generally, to using higher-level knowledge and strategies to regulate lower-level performance. Previous research suggests that metacognition is an important part of learning among adults (Baker, 1989; Garner & Alexander, 1989; Pressley & Ghatala, 1990) and children (Alexander, Carr, & Schwanenflugel, 1995; Borkowski & Muthukrishna, 1992). Metacognition contributes to learning in several ways, but especially by helping learners to use their attentional resources more efficiently, to process information at a deeper level, and to monitor their performance more accurately.

Notwithstanding its importance, there is considerable debate regarding how to measure metacognition. At the heart of the problem is the elusive nature of metacognitive knowledge itself. Most theorists assume metacognitive knowledge is highly abstract and cuts across domain-specific boundaries (Brown, 1987; Flavell, 1987; Paris & Byrnes, 1989; Schraw, Dunkle, Bendixen, & Roedel, 1995; Schraw & Moshman, 1995). In contrast, most declarative and procedural knowledge in memory is welded to a specific domain, and can be stated as a declarative fact or demonstrated through a procedure. As a result, declarative and procedural knowledge are much easier to identify, manipulate, and measure than metacognitive knowledge. Added to

this is the fact that metacognitive knowledge is acquired gradually over long periods of time, emerges relatively late in development, and often is difficult to explicate even when an individual demonstrates a high degree of metacognitive competence (Brown, 1987; Garner, 1994; Weinert & Kluwe, 1987).

Another problem is that metacognitive processes such as planning and evaluation are difficult to measure directly.

For this reason, researchers have relied on a variety of indirect measures such as verbal reports, think-alouds, self-report inventories, and subjective measures of performance accuracy. One consequence of the unobservable nature of metacognitive knowledge and regulation is that researchers have focused their attention on several specific aspects of metacognition that are easier to measure than others, especially various forms of monitoring. Most studies have focused on memory monitoring (Cavanaugh & Perlmutter, 1982; Johnson, Hastroudi, & Lindsay, 1994; Lovelace, 1984; Koriat, 1993; Schneider & Pressley, 1989), comprehension monitoring (Glenberg & Epstein, 1985; Leonesio & Nelson, 1990; Weaver, 1990), or performance monitoring (Glenberg, Sanocki, Epstein, & Morris, 1987; Pressley & Ghatala, 1990).

This chapter addresses problems related to the measurement of metacognition in greater detail. We believe that some of the more imposing obstacles can be addressed successfully via computer-based testing procedures, but especially those pertaining to the assessment of metacognitive control processes. We will argue that computer-based testing provides opportunities for researchers to measure control processes with much greater precision than with noncomputerized methodologies. Computer-based testing enables us to do so in an unobtrusive, reliable manner that is less apt to be confounded by preexperimental knowledge and ability.

The remainder of this chapter is divided into six sections. The first of these provides a brief overview of previous research and presents a multilevel model of metacognition that distinguishes between two major components, including knowledge about and regulation of cognitive processes and knowledge. We further distinguish between two subcomponents of metacognitive regulation, including metacognitive control and monitoring. Control processes are used to select performance goals and guide ongoing cognitive activities. Monitoring processes are used to evaluate the present success of one's performance and the degree to which one has met one's long-term performance goals. We assume that control and monitoring are reciprocally linked in a manner that facilitates self-regulation during performance.

The overview is followed by a section that outlines some of the methodological shortcomings of previous research. These include issues pertaining to the reliability and construct validity of dependent variables used in these studies. Of greater importance, this section considers how dependence on a limited repertoire of methodological strategies has precluded inquiry along two important lines. The first concerns the investigation of metacognitive control. We believe that few studies have investigated control processes at all, and that none have done so directly. The second line of inquiry concerns the relationship between control and monitoring processes. Current conceptualizations of metacognition make a number of assumptions about this relationship that have not been tested empirically.

The next section provides a review of recent developments in computer-based testing that offer great promise for the measurement of metacognition. These include the contribution of item-response theory to the rapidly growing field of computerized adaptive testing (i.e., tests in which a computer-controlled algorithm selects test items from a multilevel, calibrated item pool) and self-adapted testing (i.e., tests in which examinees select item of a designated difficulty level from a multilevel, calibrated item pool).

We consider ways that self-adapted testing (SAT) can be applied to the measurement of metacognition in the next section. This includes some of the psychometric advantages of SAT as well as a description of on-line measures of cognitive and metacognitive behavior that can be used to test the model of metacognition proposed later in this chapter. Specifically, we address how SAT can be used to assess metacognitive control in a variety of ways, including a measure of how accurately individuals select test items, as well as selection times, item response times, and across-test item selection strategies.

The final section outlines an agenda for future research using SAT. One important goal of this research is to link the kinds of data collected in previous studies with the kind of on-line measures available in SAT. Among other things, this would enable researchers to compare the reliability of subjective paper-and-pencil judgments made before, during, or after testing to objective measures collected during SAT. Ideally, one would hope for a strong correspondence between the two; however, one possibility is that pre- and post-test subjective judgments do not correspond closely to actual on-line item selection strategies. Another goal is that researchers investigate in detail the relationship between control and monitoring. One would expect these processes to be linked reciprocally, even though there is no direct empirical evidence to support this assumption. Establishing

such a relationship would suggest that control and monitoring processes are part of a larger regulatory system. In contrast, finding that the two processes are not related strongly would suggest that each is governed by a separate reservoir of knowledge.

The final section of the paper summarizes our main points and offers some general conclusions. Chief among these is the claim that researchers may benefit by incorporating recent innovations from the computer-based testing community, and by using SAT to bridge the gap between existing metacognitive theory and empirical studies that do not adequately address questions raised by this theory.

COGNITIVE AND METACOGNITIVE PROCESSES

Individuals rely on both cognitive and metacognitive skills when learning (Garner & Alexander, 1989; Pressley, Borkowski, & Schneider, 1987). Cognitive skills are those that help a person perform a task; metacognitive skills are those that help a person regulate and monitor task performance (Artzt & Armour-Thomas, 1992; Schraw, 1994; Slife & Weaver, 1992). Metacognition is thought to include two main components (Baker & Brown, 1984; Brown, 1987; Jacobs & Paris, 1987). The first, *knowledge of cognition*, refers to what individuals know about their own cognition or about cognition in general. It usually includes three different kinds of metacognitive awareness: declarative, procedural, and conditional knowledge (Brown, 1987; Jacobs & Paris, 1987; Schraw & Moshman, 1995). Declarative knowledge refers to knowing "about" things. Procedural knowledge refers to knowing "how" to do things. Conditional knowledge refers to knowing the "why" and "when" aspects of cognition. The second, *regulation of cognition*, refers to metacognitive activities that help control and monitor one's learning. Although a number of regulatory skills have been described in the literature (Jacobs & Paris, 1987; Kluwe, 1987), two that appear to be essential are control and monitoring processes.

A growing number of studies have been conducted over the past decade investigating these components. Those focusing on the knowledge of cognition component typically employed either think-aloud (Swanson, 1990) or self-report measures (Dixon, Hultsch, & Hertzog, 1988; Schraw & Dennison, 1994). Those focusing on the regulation of cognition component, but especially the monitoring subcomponent, typically employed some form of priming task, or asked individuals to make subjective judgments of confidence, ease of comprehension, or overall learning prior to or subsequent to completing a test

(Glenberg, Sanocki, Epstein, & Morris, 1987; Leonesio & Nelson, 1990; Weaver, 1990).

Unfortunately, because many of these studies used widely different materials, data collection procedures, and criterion measures, results are mixed and often difficult to compare. In lieu of a comprehensive review of these diverse findings, we turn briefly to a summary of recent research investigating the control and monitoring subcomponents. We do so for two reasons. One is to provide a more detailed definition of each construct. A second is to delineate the strengths and weaknesses of recent empirical research.

Research on Control

Metacognitive *control* refers to regulatory processes that occur prior to or during a learning activity that direct the course of cognitive activities. These processes include but are not restricted to planning, allocating resources, selecting strategies, and setting specific performance goals. Control processes typically are assumed to guide cognitive activities in a top-down manner (Nelson & Narens, 1990, 1994). Most theorists also assume that control processes are intentional, nonautomated, and partially statable (Bjorklund & Harnishfeger, 1990; Pressley, Harris, & Marks, 1992).

Many studies have investigated the effect of strategy instruction on metacognitive control (see Garner, 1987; Pressley et al., 1987; Pressley, 1995, for reviews). These studies invariably indicate that strategy instruction increases metacognitive control in two ways: through better use of limited cognitive resources and more elaborative processing (Willoughby, Wood, & Khan, 1994; Wood, Pressley, & Winne, 1990). However, few of these studies have shown attempts to assess the accuracy of enhanced control processes, the degree to which learners have metacognitive awareness about enhanced control, and the extent to which enhanced control is related to monitoring accuracy.

Several studies have investigated the relationship between control and monitoring more directly. Pressley and colleagues (see Pressley & Ghatala, 1990, for a review) found that experimental manipulations that improved performance (presumably by enhancing metacognitive control) did not lead to more accurate monitoring among college students. In contrast, Maki and colleagues (Maki & Serra, 1992; Maki, Foley, Kajer, Thompson, & Willert, 1990) found that experimental manipulations that neces-

sitated deeper information processing (e.g., asking readers to generate missing text information) led to more accurate monitoring.

Other studies have used estimates of future performance on a specific task as a measure of metacognitive control. Schraw (1994) asked college students to estimate their ability to monitor accurately their reading comprehension. Control predictions were correlated positively (i.e., $p < .01$) with test performance and post-test estimates of monitoring accuracy. Levels of self-assessed monitoring ability also were related to item-by-item and end-of-test monitoring accuracy. Those who rated themselves as normatively accurate monitors tended to be more accurate and to improve more than poor monitors as a function of self-generated feedback. These findings suggested that older learners possess knowledge about metacognitive processes and use this knowledge strategically to control their performance and monitoring.

A follow-up study by Schraw (1995) examined performance control judgments (i.e., pretest estimates of one's ability to perform well in a specific domain) across a variety of content domains and test formats. Results indicated that control judgments were correlated positively among domains even when test performance was controlled statistically. This suggested that metacognitive control may be a domain-general rather than domain-specific phenomenon. However, control judgments across different types of tests (i.e., recognition of facts versus recall of inferential relationships) were unrelated. This suggested that control judgments may be dependent on the specific cognitive processes required of a particular test format (see Pressley & Ghatala, 1990, and Schwartz & Metcalfe, 1994, for a further discussion).

Research on Monitoring

Metacognitive *monitoring* refers to processes that occur during or after a learning activity that provide information about the effectiveness of those activities. These processes are used to evaluate the present success of one's performance and the degree to which one has met one's long-term performance goals. Monitoring is important because it provides self-generated feedback to the control system. Without accurate monitoring, efficient control of one's performance may be impossible. Most theorists assume that monitoring is a data-driven process; that is, monitoring accuracy may be a function of domain familiarity, automaticity, and task difficulty (Koriat, 1993; Nelson & Narens, 1990, 1994).

Monitoring studies typically require individuals to make subjective judgments of learning or test performance during or after an initial study phase. Judgments are made for each test item using a 5- or 7-point Likert scale, although some researchers have used other techniques such as a continuous, bipolar scale adapted from the multidimensional literature (see Schraw, Potenza, & Nebelsick-Gullet, 1993, for a further description). The main purpose of these studies is to determine the degree to which individuals accurately assess their learning and performance.

Four types of judgments have been used in the adult monitoring literature, including *ease of learning* (i.e., judgments of encoding difficulty), *judgments of learning* (i.e., the degree to which information was learned during the study phase), *feeling of knowing* (i.e., the degree to which one has access to previously learned information in memory), and *performance judgments* (i.e., assessments of performance accuracy). These four types of judgments have been used by researchers to operationalize metacognitive processes involved in the acquisition, retention, and retrieval of information (Nelson & Narens, 1994).

Monitoring studies differ widely with respect to the type of criterion measure used to assess monitoring ability. Many studies use some form of correlation, although a number of studies report other measures such as bias (Schraw & Roedel, 1994), accuracy (Tobias, 1996), discrimination (Lundeberg, Fox, & Puncochar, 1994), or a multicomponent measure based on bias, correlation, and discrimination (Yates, 1990). Currently, there is widespread disagreement about the relative effectiveness of these measures (Keren, 1991; Liberman & Tversky, 1993; Nelson, 1984; Schraw, 1995). One point of agreement is that different criterion measures affect both observed results and how researchers interpret these results.

These studies generally suggest that adults monitor their learning and performance with a moderate degree of success, although results vary from study to study. Surprisingly, monitoring proficiency does not appear to be related strongly to relevant domain knowledge (Glenberg & Epstein, 1987; Morris, 1990; Schraw et al., 1995) or academic achievement (Pressley & Ghatala, 1988, 1990). These conclusions have been supported in the children's monitoring literature as well, although there is considerable debate regarding whether children monitor as accurately as adults (Alexander, Carr, & Schwanenflugel, 1995; Butterfield, Nelson, & Peck, 1988).

Situational constraints also affect estimates of monitoring proficiency. One constraint is the point in the learning-test sequence in which monitoring judgments are made. A number of studies indicate

that calibration of comprehension (i.e., the correlation between pretest judgments and actual test performance) is often quite poor, with most studies reporting correlations in the .00 to .25 range (Glenberg et al., 1987; Pressley & Ghatala, 1990). In contrast, calibration of performance (i.e., the correlation between posttest judgments and actual test performance) appears to be much better in both children and adults, often ranging from .30 to .50 (Glenberg et al., 1987; Maki & Serra, 1992; Maki et al., 1990; Pressley & Ghatala, 1990).

A second constraint is that specific testing conditions affect monitoring proficiency. For example, calibration of comprehension can be improved under the following circumstances: (a) when adjunct questions similar to post-test questions are provided during study (Pressley, Snyder, Levin, Murray, & Ghatala, 1987), (b) when periodic feedback is provided to test takers (Ghatala, Levin, Foorman, & Pressley, 1989), (c) when expert knowledge about the to-be-learned material is *minimized* (Glenberg & Epstein, 1987), and (d) when test takers generated missing text information (Maki et al., 1990). Surprisingly, calibration of comprehension does not appear to improve when learners were specifically requested to monitor their comprehension or when they are given the opportunity to re-study the to-be-learned materials (Ghatala et al., 1989), or when they were given practice questions prior to study (Maki & Serra, 1992).

Like calibration of comprehension, calibration of performance improved under a number of testing conditions, especially when adjunct questions were provided during the study phase (Pressley et al., 1988), when test takers received external incentives to improve monitoring accuracy (Schraw et al., 1994), and when test takers received recall rather than recognition tests (Pressley, Ghatala, Woloshyn, & Pirie, 1990). Calibration of performance also was related to level of test performance (Schraw & Roedel, 1994). Individuals monitored with less bias when judging their performance on easy rather than more difficult items.

A third general constraint is that monitoring proficiency improves with feedback, incentives, practice, and training. Stock, Kulhavy, Pridemore, and Krug (1992) found that experimenter-provided feedback increased the accuracy of confidence judgments. Schraw (1994) reported that pre-experimental estimates of monitoring proficiency were related to both local (i.e., the accuracy of item-specific performance judgments made during testing) and global (i.e., judgments of overall performance made after testing) monitoring accuracy. The accuracy of local monitoring was correlated positively to the accuracy of global monitoring. In addition, the change in

monitoring accuracy between local and global monitoring improved significantly among good monitors, but did not improve among poor monitors.

Monitoring proficiency also improves when individuals are given incentives to monitor their performance more accurately. Schraw et al. (1993) found that additional course credit for normatively high monitoring accuracy led to more accurate monitoring, whereas additional credit for normatively high test performance had no effect on monitoring accuracy. In addition, incentives to monitor more accurately improved test performance even though incentives to perform better did not.

Monitoring training also improves performance. Delclos and Harrington (1991) examined fifth and sixth grader's ability to solve computer problems after assignment to one of three conditions. The first group received specific problem-solving training, the second received problem-solving plus self-monitoring training and practice, and the third received no training. The monitored problem-solving group solved more of the difficult problems than either of the remaining groups and took less time to do so. The group receiving problem-solving and monitoring training also solved complex problems faster than the control group.

Summary

The control and monitoring research summarized above leads to a number of conclusions. Regarding control, most adults achieve some degree of metacognitive control by using helpful learning strategies. Second, many adults possess some explicit metacognitive knowledge about their ability to control performance. Third, metacognitive control in one domain tends to be related to control in another domain, even when performance is taken into consideration. Fourth, metacognitive control appears to be superior in adults (Alexander et al., 1995).

Regarding monitoring, adults monitor their performance with a moderate degree of accuracy. Monitoring improves as tests become easier and more factual. Second, monitoring proficiency appears to be independent of intellectual ability (Alexander et al., 1995; Swanson, 1990) and academic achievement (Pressley & Ghatala, 1988). Third, monitoring proficiency may be independent or even negatively related to domain knowledge (Glenberg & Epstein, 1987), independent of ease of comprehension judgments (Leonesio & Nelson, 1990), but correlated with other types of metacognitive knowledge (Schraw,

1994; Schraw & Dennison, 1994). Fourth, one's ability to monitor one's performance may improve with practice (Delclos & Harrington, 1991).

PROBLEMS WITH CURRENT MEASUREMENT APPROACHES

It could be argued that the gap between metacognitive theory and empirical research is as great as any other area of psychological inquiry. These are several specific reasons for this state of affairs, many of them being methodological in nature (Kruglanski, 1989). This section divides these problems into three interrelated categories that are ranked ordered from our vantage point in order of importance. The three categories include *task*, *test*, and *person* constraints on the measurement of control and monitoring.

Task Constraints

Task constraints refer to characteristics of the experimental task that impede measurement of either control or monitoring processes. The most serious obstacle is that researchers cannot manipulate either control or monitoring processes directly, but must be content to manipulate the task environment in which control and monitoring are performed. This means that researchers must make inferences about complex metacognitive processes on the basis of indirect measures. Although this is certainly not a new problem to psychologists, it is a serious one.

Operationalizing metacognitive control has been an especially virulent problem. Presumably, the best way to study control processes would be to allow the examinee to exercise a great deal of strategic control over his or her performance. Previous studies have attempted to do so by providing specific task information, learning goals, opportunity to study, strategies for learning, or conditions under which learning is facilitated. In essence, these studies examined whether a variety of experimental factors affected metacognitive control. However, none of these studies allowed examinees to demonstrate overtly in a directly observable manner how they attempted to control their test-taking behavior. One way to do so would be through the use of on-line verbal protocols in which individuals describe their cognitive processes (Ericsson & Simon, 1984; Pressley & Afflerbach, 1995). However, although an important research tool, verbal reports are intrusive, resource consuming, and assume that individuals have explicit access to metacognitive processes.

An alternative would be to study the way examinees make strategic choices throughout a test. In self-adapted testing, for ex-

ample, individuals choose test items of a designated difficulty level from a multilevel, calibrated item pool. This may enable researchers to examine several aspects of metacognitive control in an explicit, yet unobtrusive manner. One aspect is the goodness of fit (i.e., calibration accuracy) between self-selected items and observed performance. Another aspect is whether examinees show evidence of improved accuracy over the course of the entire test.

A somewhat different task constraint is introduced when researchers ask examinees to make subjective judgments of learning and performance while simultaneously performing complex tasks. Researchers invariably assume that such ratings have little effect on performance, although oddly, there are no empirical studies we know of that have investigated this assumption. Of greater importance, researchers also assume that the demands of taking a test have little impact on the accuracy of subjective ratings. This assumption clearly is untenable in that confidence judgments become increasingly more biased as a function of test difficulty (Schraw & Roedel, 1994; Schwartz & Metcalfe, 1994). Although researchers have attempted to compensate for such problems via the judicious use of statistical analyses (cf. Nelson, 1984), no amount of statistical tinkering can eliminate these problems entirely (cf. Funder, 1987; Keren, 1991; Liberman & Tversky, 1993; Schraw, 1995).

Test Constraints

Test constraints refer to characteristics of the test itself, rather than the test environment, that impede measurement of either control or monitoring processes. A recent review by Schwartz and Metcalfe (1994) addressed four test-related problems that we summarize here. One source of variation among examinees, and presumably an important source of measurement error, pertains to the type of test being given. Recall tests often are assumed to be more cognitively demanding than recognition tests. Most empirical studies echo this difference by revealing higher correlations between performance and confidence (or accuracy) judgments on recall tests. One reason for higher correlations is less restriction of the range of scores on recall tests when compared to recognition tests. Because recall tests are more difficult, their scores will vary across a wider range of possible values. In contrast, easier recognition tests restrict the observed range of a correlation due to homogeneous performance or ceiling effects.

Another inadvertent problem of recognition tests is that examinees are influenced by the availability of information included in the

test item. Because test answers are provided explicitly in a recognition test, but must be generated in a recall test, examinees are significantly more confident when monitoring recognition tests, but more accurate when monitoring recall tests (Ghatala, Levin, Foorman, & Pressley, 1989).

A second major source of measurement error is the length of a test, or if it is a recognition test, the number of alternatives from which one may choose for each item. It is well established that a test's reliability is directly related to its length, with longer tests, and recognition tests with more alternatives, being more reliable (Crocker & Algina, 1986). Unfortunately, many early studies of monitoring used multiple tests with one or two items per test, rather than the preferable one test with a large number of items. To illustrate, Glenberg et al. (1987) reported no statistically significant relationship between pretest judgments of learning and subsequent performance. This group of experiments required individuals to answer one main idea question per test for a large number of tests. Replicating this study, having first increased the length of each test, Weaver (1990) found that the observed value of r increased monotonically as a function of test length, until it reached an asymptotic value of $r = .60$. Thus, Glenberg et al. (1987) failed to identify a significant relationship between judgments of learning and test performance due to unreliable test scores.

A third source of error is test difficulty. Monitoring accuracy declines as a test becomes more difficult, even when test performance is controlled statistically (Schraw & Roedel, 1994; Schraw, Dunkle, Bendixen, & Roedel, 1995). In addition, overconfidence is more common than underconfidence and more likely to occur when a test is difficult (Cutler & Wolfe, 1989; Newman, 1984). These patterns have been observed on a variety of tasks including probability judgments (Fischhoff, 1988), reading comprehension (Glenberg et al., 1987), recalling emotions (Thomas & Diener, 1990), and social judgments (Dunning, Griffin, Milojkovic, & Ross, 1990).

There are at least two reasons a difficult test may interfere with control and monitoring processes. One is that individuals lack sufficient background knowledge to answer the test question. It is well known that individuals resort to a number of helpful, but fallible, heuristics under these circumstances that bias their judgments (Fischhoff, 1988; Tversky & Kahneman, 1973). A second reason is that information in memory is inaccessible during testing (i.e., available in memory, but presently unretrievable). Partial or total inaccessibility may lead to severe judgment bias due not only to poor monitoring, but fallible retrieval processes as well (Koriat, 1993, 1994).

A fourth source of error is knowledge about the test. Test-relevant knowledge may affect control and monitoring in several ways—namely, by enabling examinees to identify test-relevant information more efficiently, process information at a test-appropriate level (McDaniel & Einstein, 1989), and utilize self-generated feedback (Glenberg et al., 1987). In general, as knowledge of the test increases, performance and the reliability of tests improve as well (Schwartz & Metcalfe, 1994). Research by Metcalfe (1993) also found that administering a test that was not expected reduced the correlation between performance judgments and actual performance dramatically.

Person Constraints

There are a number of ways that prior knowledge might affect control and monitoring processes negatively, and thereby reduce the reliability of measurements (Baker, 1989; Garner & Alexander, 1989). Insufficient knowledge may preclude the use of helpful learning and test strategies and lead to lower performance. Lower performance may, in turn, lead to a restriction in the range of observed test scores. Low domain knowledge also makes a test more difficult, which has several deleterious effects on monitoring already described above.

It is possible that prior knowledge interacts with many of the constraints described above in complex ways. For example, low prior knowledge presumably affects the degree to which individuals learn information during a pretest study session. Poorer learning leads to a greater amount of inaccessible information and a more difficult test. Low prior knowledge in a domain also may restrict deeper information processing that could affect performance on some test questions, but not others.

It is important to note, however, that increasing prior knowledge per se does not seem to improve monitoring (Nelson & Narens, 1990; Pressley et al., 1990), unless the inclusion of prior knowledge provides an opportunity for self-generated feedback or additional knowledge about the test itself (Glenberg et al., 1987). For example, research by Morris (1990) found that although knowledge was related positively to performance, it was not related to monitoring accuracy. Schraw (Schraw & Roedel, 1994; Schraw et al., 1995) extended these findings across multiple domains, arguing that individuals possess a domain-general (i.e., knowledge-independent) monitoring skill that is independent of domain knowledge. Glenberg and Epstein (1987) also reported that higher levels of expert knowledge actually interfered with accurate monitoring.

Summary

Empirical studies of control and monitoring lag behind metacognitive theory. One important reason is that each of these processes is difficult to operationalize experimentally and to manipulate directly. Researchers have relied on several limited measurement paradigms, including error detection (see Baker & Cerro, this volume) and subjective calibration judgments. Both of these methodologies are fraught with measurement problems related to the nature of the task itself, to factors including the type and difficulty of the test, and to characteristics of the examinee.

In subsequent sections of this chapter we argue that self-adapted testing allows researchers to eliminate many of these problems, and thereby increase the construct validity of tests (Rocklin, O'Donnell, & Holst, 1995), by (a) controlling for test and item difficulty using a calibrated pool of independent test items, (b) reducing measurement error attributable to characteristics of the examinees such as ability and prior knowledge, (c) utilizing unobtrusive measures that do not compete for the examinees' limited resources, and most importantly, (d) allowing the test taker to exercise a much greater degree of control during the testing process. We turn now to a brief overview of computer-based testing and two recent developments: computerized adaptive and self-adapted testing.

NEW DEVELOPMENTS IN EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Item Response Theory

During the past few decades, Item Response Theory (IRT), has emerged as the psychometric model used by an increasing number of testing programs in education and psychology. For large-scale achievement and proficiency tests in particular, IRT has largely supplanted classical test theory as the basis for test development, scoring, and equating. The central concept of IRT is the item characteristic curve (ICC), which specifies the relationship between the level of an examinee's *proficiency* (i.e., estimated ability) and the probability that he or she passes the item.

The most commonly used IRT models assume that there is a monotonic relationship between examinee proficiency and the probability of passing an item. In addition, it is assumed that the set of test items under consideration is unidimensional (i.e., measures a single, unobservable construct). It has been typically found, however, that

the IRT model will adequately fit the test data if there is one sufficiently dominant factor underlying the items. A detailed explanation of IRT is beyond the scope of this chapter; the interested reader is referred to Hambleton, Swaminathan, and Rogers (1991) for a good overview of basic IRT concepts.

Two principles of IRT are particularly relevant to the present discussion. The first is a key property of IRT, termed invariance, which states that an examinee's proficiency is independent of the characteristics of the items that are administered. Consider the case in which there is an available pool of 400 test items, and that it is larger than would be administered to a given examinee (e.g., 100 items) during a testing session. Regardless of which 100 items were administered from the pool, the examinee's expected proficiency estimate would be invariant. Invariance holds because IRT-based proficiency estimates take into account both (a) characteristics (primarily difficulty) of the items that were administered and (b) the examinee's performance on those items. An important implication of invariance is that two examinees can receive completely different tests, drawn from the same item pool, yet their proficiency estimates can be compared. Any differences in difficulty of the two tests are taken into account by the IRT estimation procedure.

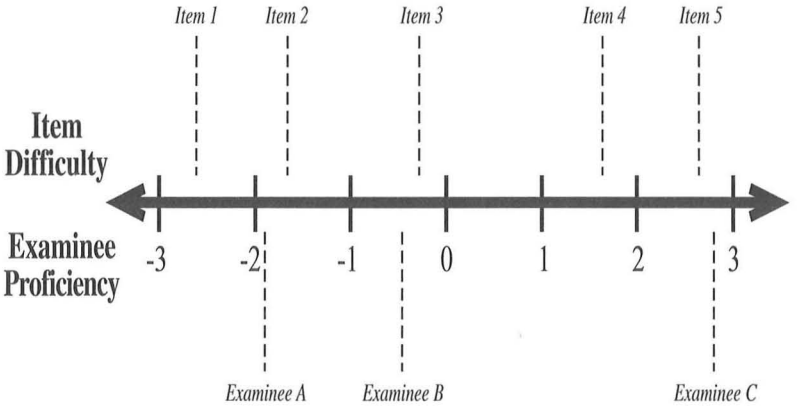
It should be noted that invariance is not a feature of the classical test theory measurement model, in which proficiency estimation is based solely on test performance (i.e., number of items passed). If two examinees take two tests that differ in difficulty, then the difference between the examinees' proficiency levels is confounded with the difference between the difficulty of the tests.

A second principle of IRT that is of particular relevance to the study of examinee monitoring and control is that the difficulty parameters of the ICCs, which indicate the relative difficulties of the items, are placed on the same scale as examinee proficiency. This joint scaling is depicted in Figure 1, which indicates that Item 1 is the least difficult item, followed by Item 2, and so on through Item 5. Moreover, examinee A is the least proficient of the three examinees, and Examinee C the most proficient.

Measuring difficulty and proficiency on the same scale allows one to assess the degree of match between the difficulty of an item and an examinee's proficiency. Why is this joint scaling important? The closer the match between an item's difficulty and an examinee's proficiency, the more informative is the examinee's response to that item in estimating his/her proficiency. Hence, more difficult items are most informative for more proficient examinees, whereas less

difficult items are most informative for less proficient examinees. In Figure 1, the most informative items for Examinees A, B, and C are Items 2, 3, and 5, respectively.

Figure 1. The joint scaling of item difficulty and examinee proficiency in IRT.



Computer-Based Testing

With the introduction and rapid proliferation of microcomputers came an increased use of computers to administer tests. There are a number of advantages realized with computer-based testing that may make it attractive to examinees, including the capability for on-demand testing, as well as immediate test scoring and reporting of results. From a researcher's standpoint, however, computer-based testing provides two additional advantages. First, it allows a much greater degree of control over the test administration. Such control may include (a) the order in which items are considered and answered, (b) how long each item is presented, and (c) whether or not examinees are allowed to review, and possibly change, their answers to items. Second, it allows the researcher to unobtrusively collect a great deal of information about the test session, such as how long an examinee took to respond to each item or whether or not the examinee changed his/her answers to any test items. Because of these advantages, a computer-based test provides a unique opportunity for re-

searchers to study examinee test-taking behavior. With paper-and-pencil tests, such advantages are unavailable.

Computerized Adaptive Testing. Computerized adaptive testing (CAT) combines the psychometric advantages of IRT with the computing power of current microcomputers. In a CAT, a computer algorithm is used to match the difficulty of the items administered to the estimated proficiency of each examinee. At each step in a CAT, the next item to be administered is a function of the examinee's responses to previously administered items. Using a CAT, examinee ability is estimated more efficiently than with a conventional test because typically fewer items are required to attain the same degree of measurement precision. It has typically has been found that a CAT requires about half as many items to estimate an examinee's proficiency with the same degree of precision as a paper-and-pencil test.

Note that both of the IRT principles discussed earlier are essential to a CAT. Because item difficulty and examinee proficiency are on the same scale, items having difficulties matching an examinee's current proficiency estimate can readily be identified and administered. And, because examinees receive unique tests, the invariance property allows their proficiency estimates to be compared.

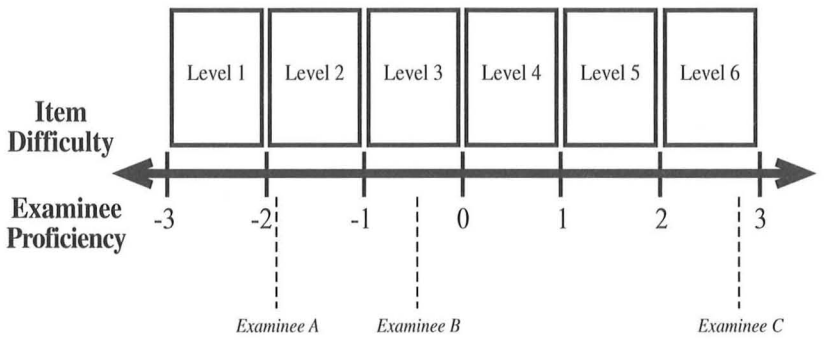
Self-Adapted Testing. Although CAT is by far the most popular application of IRT in computer-based testing, other applications have been studied. One of these is self-adapted testing (Rocklin & O'Donnell, 1987). A SAT is similar to a CAT with one important exception. In a self-adapted test, the examinee is allowed to choose the difficulty level of each test item administered, whereas in a CAT a computer algorithm chooses each item to be administered based on the examinee's performance on items administered earlier in the testing session.

In a SAT, an examinee chooses the difficulty level of each item administered from an item pool has been divided into several (typically 5–8) ordered difficulty levels, or strata, based on the IRT difficulty parameters of the items. This relationship among difficulty levels is illustrated in Figure 2. Testing begins with the examinee choosing the difficulty level of the first item, at which point an item from the chosen stratum is drawn (without replacement) in a random fashion and administered. After this item is answered, the examinee is then asked to choose the difficulty level of the next item. This procedure continues until a predetermined number of items has been administered or a desired precision of proficiency estimation has been reached. After item administration is completed, the examinee's test performance is calculated using an IRT-based proficiency estimation

method. As with a CAT, because proficiency estimation is IRT based, the invariance property insures that the test performances of different examinees receiving a SAT can be directly compared even though they may have chosen to be administered tests that varied substantially in difficulty. Successful implementation of a SAT is largely dependent on the instructions presented at the beginning of the test. It must be explained to examinees that their test performance will be evaluated on the basis of the difficulty levels they choose as well as the number of items that they pass. Because most examinees are used to taking tests where performance is based solely on how many items are passed, examinees taking a SAT may tend to choose low difficulty levels unless adequate instructions are provided. Hence it is very important to provide examinees with clear instructions when administering a SAT. An example of instructions used with a SAT are found in Wise, Plake, Johnson, and Roos (1992).

The research on SAT conducted thus far has focused on its effects on test performance and its relationship to examinee affective variables. Several studies have compared SAT with CAT, finding that examinees receiving a SAT obtained significantly higher mean proficiency estimates (Roos, Plake, & Wise, 1992; Wise et al., 1992; Vispoel & Coffman, 1994). Moreover, the difference in mean estimated proficiency between SAT and CAT has been found to interact with other variables. Significant interactions have been found between test type and examinee scores on the Test Anxiety Inventory (Spielberger, 1980), with the difference in mean estimated proficiency between SAT and CAT increasing with examinee test anxiety (Rocklin & O'Donnell, 1991; Vispoel & Coffman, 1994; Vispoel, Rocklin, & Wang, 1994; Vispoel, Wang, de la Torre, Bleiler, & Dings, 1992). In addition,

Figure 2. Item Difficulty level strata in self-adapted testing.



Vispoel et al. (1994) found a significant interaction between examinee verbal self-concept and test type, with the largest difference in mean estimated proficiency between SAT and CAT being associated with low examinee verbal self-concept.

There also is evidence that the use of a SAT moderates the relationship between examinee anxiety and test performance. In two studies comparing SAT and CAT it was found that examinees administered a SAT reported significantly lower post-test state anxiety than examinees administered a CAT (Roos et al., 1992; Wise et al., 1992). It has also been found that a SAT yields proficiency estimates that are less related to test anxiety than those obtained when a CAT or a conventional test is used (Rocklin & O'Donnell, 1991; Vispoel & Coffman, 1994; Vispoel et al., 1994; Vispoel et al., 1992). The findings from these studies suggest that use of a SAT reduces the influence of anxiety on test performance.

CONTROL, MONITORING, AND SELF-ADAPTED TESTING

Although previous research on SAT has focused on its effects on anxiety and test performance, a SAT also affords an opportunity to measure elements of metacognition. To understand this, it is useful to consider the activities of the examinee during his/her test. A difficulty level is chosen by the examinee, an item is administered, the examinee answers the item, and the examinee is provided a choice for the difficulty level of the next item. This sequence is repeated until the test is completed.

We have observed that most examinees vary their difficulty level choices during the course of a SAT. Moreover, it has been found that many examinees tend to adjust their difficulty level choices to receive items that are well-matched to their proficiency levels (Wise et al., 1992). That is, many examinees taking a SAT appear to be motivated to attain the same difficulty-proficiency match that is explicitly sought by the computer algorithm in a CAT.

What psychological processes might be involved in attaining this match? We contend that two key processes are monitoring and control. Monitoring is required to assess the difficulty of the previous item, and to compare its difficulty to one's perceived proficiency. Control is then required to make a strategic choice, regarding the next item's difficulty, on the basis of the perceived degree of match between item difficulty and proficiency. If the match is sufficiently close, then the examinee will likely choose the same difficulty level as the previous item. If the match is not judged to be close then the examinee will change difficulty levels in order to attain a closer

difficulty-proficiency match. For example, if the examinee's monitoring process yields a judgement that the previous item was too easy, then the control process will choose a more difficult next item.

Thus, whereas most of the previous research on SAT has focused on the outcomes of taking a SAT, there is important information to be gained by studying the *process* of taking a SAT. Through an analysis of the SAT experience, we see that, although both monitoring and control play a major role in the examinee's strategic choices, the observable examinee behavior (difficulty level choice) most directly reflects the control process. Later in this chapter we outline several ways of using the data from a SAT to construct measures of the control process.

Some Methodological Advantages of SAT

Self-adapted testing provides a unique, unobtrusive method for gathering information about metacognitive processes, and especially of control. Indeed, examinees need not be given specific instructions about control or monitoring, or even know that their test behaviors provide relevant information about these processes. The fact that control processes are studied unobtrusively has two important advantages. One is that examinees are able to focus all of their resources on the test, rather than dividing their attention between performance and control-assessment activities. A second advantage is that direct measures of metacognitive control are available (i.e., item selection time and accuracy), rather than an indirect, subjective assessment of control (i.e., confidence or accuracy judgments).

SAT has a number of other advantages as well that pertain specifically to the task, test, and person constraints described earlier in this chapter. The most important of these is examinee control. Whereas all previous studies have asked examinees to complete a test designed by researchers, SAT enables an examinee to choose items that he or she feels are best suited to his or her proficiency without compromising comparability among examinees. With respect to the study of metacognition, individuals with a high degree of metacognitive control should be able to select difficult, yet answerable items. Those with less control may select test items that are less appropriate for them. Those with poor control may regularly select items that are too easy or too difficult. The self-controlled nature of SAT enables researchers to study the relationship among selection time, accuracy, and overall test proficiency, as well as a variety of self-report judgments made prior to, during, or subsequent to the test. Experiments could be expanded to examine motivational variables as well.

SAT also may increase the construct validity of proficiency estimates, and presumably measures of metacognitive control, by reducing confounds due to anxiety (Rocklin et al., 1995; Wise, 1994) and test difficulty. This helps to reduce or eliminate many of the test-based constraints typical of previous studies. For example, given that individuals select test items from a pool of calibrated items, the difficulty of these items should have little effect on the accuracy of metacognitive control. This is in stark contrast to traditional paper-and-pencil tests in which examinees monitor their performance with greater bias as test items increase in difficulty.

Another strength of SAT is the property of invariance, which enables each examinee to select items that are optimally suited to his or her proficiency. Differences in the absolute difficulty of items need not compromise estimates of metacognitive control. This means that measures of metacognitive control are comparable on the same scale even though individuals may be administered different test items and even though individuals differ with respect to underlying ability.

The fact that SAT yields comparable estimates of proficiency and metacognitive control regardless of differences in ability eliminates a crucial person constraint in the study of metacognitive processes. It is likely that prior knowledge also has less impact on proficiency and control estimates than it would using paper-and-pencil tests. Although prior knowledge may greatly affect which items an examinee selects, item selection in itself does not affect estimates of proficiency. On the other hand, it is possible that individuals with no prior knowledge, or a great deal of it, may be poorly suited to the test if there are an insufficient number of test items near their true proficiency level.

In summary, we believe that self-adapted testing provides a unique opportunity to study on-line metacognitive control processes in an unobtrusive manner. The ability to do so permits researchers to explore a number of theoretical relationships among control, monitoring, and other cognitive skills (e.g., working memory span) that remain unanswered. We describe several intriguing questions in a subsequent section on future research. First we describe two methodological constraints on the use of self-adapted testing, then we describe a number of direct or derived measures of metacognitive control that are available from a typical SAT testing session.

Two Methodological Considerations

Two key issues must be addressed when using a SAT to measure metacognitive control strategies. First, the distribution of the item difficulties should span the range of the distribution of examinee

proficiencies, with enough items throughout the range that an examinee could take an entire SAT consisting of items with the same general level of difficulty. Having an item pool that is both "wide" and "deep" prevents examinees from being administered items that are not well matched to proficiency solely because well-matched items are unavailable.

A second consideration concerns the instructions given to the examinees. Without instructions for examinees to try to attain a close difficulty-proficiency match, it is unclear whether examinees who did not choose closely matched items did so because they were unable to match well or because they chose their items to attain another goal (e.g., reduction of test anxiety). Hence, examinees should be explicitly told to try to attain a close match. This, however, raises a troublesome new problem—how does one word such instructions such that examinees unequivocally understand their task?

The resources required to administer a SAT pose a third restriction on its use in metacognitive research. To administer a SAT, one must have (a) an item pool that is of sufficient size and has a broad range of item difficulty, (b) IRT parameter estimates for each item, and (c) computer software for administering computer-based tests. Regarding the item pool, it is important to have a distribution of item difficulties that spans the range of examinee proficiencies, and is "deep" enough that an examinee could choose a difficulty level that reflects a close difficulty-proficiency match many times without exhausting the difficulty level and being forced to receive items that are less well-matched. As an illustration, if a researcher plans to use eight difficulty levels in administering a 20-item SAT, the item pool should contain at least 160 items. Furthermore, IRT item parameter estimation requires a sizable calibration sample. Depending on the IRT model used, the typical recommendations for minimum calibration sample size range from 200 to over 1,000 examinees. Finally, special microcomputer software is needed to administer the SAT, such as the MicroCAT Testing System (Assessment Systems, 1994). Roos, Wise, Yoes, and Rocklin (in press) describe the program code needed to administer a SAT on the MicroCAT system.

QUANTIFYING METACOGNITIVE PROCESSES USING SELF-ADAPTED TESTING

Although both monitoring and control processes appear to be at work in a self-adapted test, the control process is more easily quantified using measures obtained during the testing process. A self-adapted test that is administered using computer-based testing

software such as MicroCAT (Assessment Systems, 1994) can provide a variety of information that is relevant to the measurement of metacognitive activities. When a MicroCAT test is administered, an output file for each examinee is created containing a detailed record of the examinee's testing session. The file can contain an item-by-item record of the difficulty level chosen, whether the item was answered correctly or incorrectly, the examinee's current proficiency estimate, and its standard error, as well as the time taken both to choose the item difficulty level and respond to the administered item. This information is readily obtainable from the MicroCAT testing system and does not require extensive programming skills on the part of the researcher. A guide to developing self-adapted tests on MicroCAT is provided by Roos, Wise, Yoes, and Rocklin (in press).

It is important to note that these measures are obtained in an unobtrusive manner. This mode of data collection allows examinees to focus their attention entirely on the test, alleviating concerns regarding the effects on test performance of requesting examinees to provide self-reports of metacognition.

There are several ways to quantify the relationship between the metacognitive control process and test performance (i.e., proficiency). To further illustrate these quantifications, we will refer to Tables 1 and 2. Table 1 is an example of a testing session for an examinee with a good match between proficiency and item difficulty, whereas Table 2 provides an example of an examinee with a poor proficiency-item difficulty match. Each examinee is administered 20 items from a pool of calibrated items that are partitioned into six mutually exclusive difficulty levels. For each item administered, the difficulty level chosen is displayed in the second column where Level 1 contains the easiest items and Level 6 contains the most difficult items. The difficulty parameter of the administered item is displayed in the third column. The difficulty parameters of the items are obtained using IRT estimation methods; these parameter values are matched to the scale of examinee proficiency, which typically has a mean of zero and a standard deviation of one. The higher the item difficulty parameter value, the more difficult the item. The fourth column indicates the difference between the examinee's proficiency and the difficulty of the item. For example, the examinee in Table 1 had a final (i.e., end-of-test) proficiency estimate of -1.31, which is relatively low. The first item administered had a difficulty of -1.39, which indicates a close proficiency-difficulty match (.08). The fifth column lists the absolute value of the proficiency minus difficulty difference. The final column indicates the correctness of the examinee's answer to the item.

Table 1. Testing Session for an Examinee With a Good Match Between Proficiency and Item Difficulty (Proficiency = –1.31, Standard Error = .319)

Item	Difficulty Level Chosen	Item Difficulty Parameter	Proficiency – Difficulty Difference	Absolute Difference	Item Outcome
1	2	–1.39	0.08	0.08	Right
2	3	–1.13	–0.18	0.18	Wrong
3	2	–1.73	0.42	0.42	Right
4	2	–1.42	0.11	0.11	Wrong
5	2	–1.28	–0.03	0.03	Right
6	2	–1.30	–0.01	0.01	Wrong
7	2	–1.32	0.01	0.01	Right
8	2	–1.50	0.19	0.19	Right
9	2	–1.25	–0.06	0.06	Right
10	2	–1.39	0.08	0.08	Right
11	2	–1.77	0.46	0.46	Right
12	2	–1.67	0.36	0.36	Right
13	2	–1.30	–0.01	0.01	Wrong
14	2	–1.57	0.26	0.26	Right
15	2	–1.27	–0.04	0.04	Right
16	2	–1.64	0.33	0.33	Right
17	3	–0.73	–0.58	0.58	Wrong
18	3	–1.24	–0.07	0.07	Wrong
19	3	–1.12	–0.19	0.19	Wrong
20	3	–1.14	–0.17	0.17	Wrong
		Mean Over the Last 5 Items:	–0.14	0.27	

Table 2. Testing Session for an Examinee With a Good Match Between Proficiencyand Item Difficulty (Proficiency = 0.32, Standard Error = .620)

Item	Difficulty Level Chosen	Item Difficulty Parameter	Proficiency – Difficulty Difference	Absolute Difference	Item Outcome
1	3	–1.13	1.45	1.45	Right
2	2	–1.39	1.71	1.71	Right
3	2	–1.73	2.05	2.05	Right
4	3	–0.73	1.05	1.05	Right
5	2	–1.42	1.74	1.74	Right
6	3	–1.24	1.56	1.56	Right
7	3	–1.12	1.44	1.44	Right
8	3	–1.14	1.46	1.46	Wrong
9	2	–1.28	1.60	1.60	Right
10	2	–1.30	1.62	1.62	Right
11	2	–1.32	1.64	1.64	Right
12	2	–1.50	1.82	1.82	Right
13	2	–1.25	1.57	1.57	Right
14	2	–1.39	1.71	1.71	Right
15	3	–0.54	0.86	0.86	Right
16	2	–1.77	2.09	2.09	Right
17	2	–1.67	1.99	1.99	Right
18	2	–1.30	1.62	1.62	Right
19	2	–1.57	1.89	1.89	Right
20	2	–1.27	1.59	1.59	Right
		Mean Over the Last 5 Items:	1.84	1.84	

The first measure of the relationship between metacognitive control and test performance is provided by the proficiency-difficulty differences. If instructed to attain a close proficiency-difficulty match, examinees should proceed through the test, monitoring the difficulty of the items administered and attempting to control subsequent difficulty level choices to attain a close proficiency-difficulty match. The degree to which an examinee is successful in controlling item difficulty will be reflected by the magnitude of his/her proficiency-difficulty difference at the end of the test, with smaller differences indicating greater control. Because the items are typically arranged randomly within each difficulty level, perhaps a more reliable index of the proficiency-difficulty match is provided by the average difference taken over the last five items. This is a measure of *bias*—the degree to which an examinee tends to select items that are too easy or too difficult. For example, the examinee in Table 1 showed a very good proficiency-difficulty match (-.14), whereas the examinee in Table 2 exhibited a poorer match (1.84) indicating a bias towards choosing item difficulties that were too low.

Another measure of control is provided by the absolute value of the proficiency-difficulty difference. This is an index of *accuracy*—the degree to which selected item difficulties are matched to an examinee's proficiency. This index is also quite different for the examinees in Tables 1 and 2, with the examinee in Table 1 exhibiting a substantially more accurate match.

The standard error of the final proficiency estimate provides an alternative measure of accuracy. The more consistent the examinee is in choosing items well-matched to his/her proficiency level, the smaller the resultant standard error. Hence, the magnitude of the standard error indicates the accuracy of the examinee choices. The standard error for the examinee in Table 1 (.319) is substantially smaller than that for the examinee in Table 2 (.620).

Additional information is available from the testing session that may also prove useful in the study of control and (possibly) monitoring processes. One general type of information available is response latency; that is, the amount of time examinees take to (a) choose item difficulty levels and (b) answer items. Measures of this sort are very difficult to obtain in a traditional paper-and-pencil test but are easily and unobtrusively obtained when a test is administered via computer.

Researchers also may gain a better understanding of the control process through an investigation of the strategies used by examinees in selecting item difficulties. A computerized adaptive test provides an efficient model of control because the item selection algorithm

strives for a close proficiency-difficulty match. It would be particularly interesting, for example, to identify examinees who behave nearly as efficiently (or perhaps even more efficiently) as the computerized adaptive algorithm.

DIRECTIONS FOR FUTURE RESEARCH

Self-adapted testing allows researchers to investigate at least six questions pertaining to metacognitive control, and the relationship between control and monitoring, that have not been addressed adequately in previous research. We present these questions beginning with the most obvious and specific ones, gradually moving toward broader, more theoretical concerns.

Question one pertains to the relationship between metacognitive control and test performance. Researchers often assume that more accurate control leads to better test performance. SAT enables one to test this relationship directly while eliminating confounds due to item difficulty and presumed underlying ability. Existing theory also predicts a strong relationship between the accuracy of control judgments and performance (Nelson & Narens, 1994; Schraw, 1994). Researchers could study the impact of practice, domain familiarity, instructions, and other test-specific constraints via direct manipulation of these variables. Similarly, person-related variables such as prior knowledge and working memory span could be examined via blocking procedures, or treated as covariates.

Question two addresses the relationship between metacognitive control and response latency variables, including item selection and item response times. It is important to note that measures of response latency do not provide pure measures of a single cognitive activity per se. For example, item selection times, especially in the middle and later parts of a test, reflect some mix of control, monitoring, and performance processes. Nevertheless, SAT provides the best available methodology for assessing the relationship between control accuracy and response time. There is little theoretical precedent thus far regarding the relationship between control mechanisms and latencies. In general, response time and performance are related inversely, although the magnitude of the relationship, as well as its direction, depends on the type of variables being compared (Meyer, Irwin, Osman & Kounios, 1988). We expect a similar relationship between control accuracy and item response times. It is unclear, however, how control accuracy and item selection times are related. One plausible scenario is that individuals with a high degree of metacognitive control need little time to make strategic decisions, in part, because

many of these decisions are automated. This should lead to an inverse relationship between selection time and accuracy; that is, as control increases, selection times decrease. On the other hand, if item selection times include monitoring processes carried over from the previous item, we would expect a negative relationship between selection time and control accuracy. This assumes that monitoring is a relatively nonautomated, time-consuming process.

Data collected from SAT studies can be used to test competing hypotheses about the relationship between selection and response times, and control accuracy. One possibility is that this relationship changes systematically as a function of examinee knowledge, proficiency, practice, or test efficacy (Rocklin et al., 1995; Wise, 1994). These changes could be studied easily by blocking examinees on any of these variables or by manipulating controllable variables (e.g., instructions) directly.

Question three pertains to the specific relationship between expertise and control processes. Opinion appears to be split on this matter. Some researchers have suggested that monitoring accuracy is largely a by-product of domain-specific expertise (cf. Glaser & Chi, 1988). However, a number of recent studies (Glenberg & Epstein, 1987; Morris, 1990; Schraw & Roedel, 1994) failed to show a relationship between monitoring accuracy and domain expertise. It is important to note, however, that the relationship between monitoring and expertise may be quite different than the relationship between control processes and expertise. Currently, we know of no study that examines control accuracy across different levels of expertise.

SAT provides a format for investigating the relative impact of expertise on control processes, including performance accuracy, control accuracy, and item selection and response times. Although we would expect expertise to be positively related to test performance and estimated proficiency, we would not necessarily predict a corresponding increase in control accuracy. This reflects our view that control processes are, in part, domain-general phenomena (cf. Schraw, Dunkle, Bendixen, & Roedel, 1995). Although expertise should enable examinees to perform better on a test, their expertise need not improve their ability to control or monitor with a high degree of accuracy.

A fourth question is the degree to which control accuracy is related to other cognitive variables such as general aptitude and working memory span. Very little research has been done in this area in general. Of studies that have investigated these relationships directly or indirectly, there is little evidence that aptitude is related

strongly to metacognitive processes in children (Alexander, Carr, & Schwanenflugel, 1995; Swanson, 1990) or adults (Pressley & Ghatala, 1990; Yan, 1994). We know of no study investigating the relationship among control and monitoring accuracy and traditional indices of the speed and accuracy of working memory.

Research in this area is important for two reasons. One is to establish the degree to which metacognitive processes such as control and monitoring are related to "hard-wired" cognitive differences such as general intelligence and working memory capacity (cf. Jensen, 1992). Evidence that metacognition is not related strongly to these variables would highlight the flexible, developmental nature of metacognitive knowledge. A second reason is to examine the compensatory relationship between measures of cognitive ability and metacognitive knowledge. In a ground-breaking study by Swanson (1990), for example, metacognitive knowledge contributed to complex problem solving among young adolescents over and above the effect of ability. This finding suggests that metacognition may follow a separate developmental path, and may act independent of other cognitive mechanisms (see Alexander et al., 1995, for a further discussion).

Question five pertains to the still elusive relationship between control and monitoring processes. Several theorists have distinguished clearly between control and monitoring processes (Koriat, 1994; Nelson & Narens, 1990; Pressley & Ghatala, 1990).

Nevertheless, much of the empirical literature in the field has focused on monitoring rather than control processes, due in large part to the difficulty researchers face when measuring control.

Some believe that control and monitoring are practically, if not statistically, linked (Nelson & Narens, 1990, 1994). Others believe that monitoring is both functionally and statistically independent of control, and in fact, represents a fundamentally different type of cognitive activity (Koriat, 1993, 1994).

The literature is in need of further contributions on this point. We believe self-adapted testing methods can be used with tremendous advantage to address this question. Previously, we described how control processes may be quantified in a SAT via direct and indirect measures obtained unobtrusively. It also is possible to obtain measures of monitoring via subjective judgments made after answering a test question within the otherwise computer-based format of SAT. Control and monitoring indices could be compared over the course of a test to determine their relationship. If the two are linked, one would expect monitoring judgments made at item selection to be linked to

item selection at item $i + 1$. Data of this type, as well as a variety of derived indices of control and monitoring, could be used to test the efficacy of a *regulatory loop* that connects monitoring and control functions. In this view, monitoring processes provide data-driven feedback to control processes that use this feedback to iteratively guide future performance. This presumes that monitoring and control processes are flexible, reciprocal processes that communicate with each other, even if they do not share a common set of cognitive resources.

It is possible that control and monitoring processes are related in different ways as a function of expertise. For example, control and monitoring may be related more strongly as expertise increases, provided these processes become mutually encapsulated within the expert domain (Glaser & Chi, 1988). If control and monitoring skills remain domain-general in nature, then expertise within a domain should not matter. Another possibility is that control and monitoring are unrelated (Koriat, 1993, 1994). In this view, monitoring processes are "parasitic" in that they are based on domain knowledge and efficacy beliefs within the domain, rather than a metacognitive mechanism that actually monitors the accuracy of performance independent of domain knowledge.

A final question addresses the degree to which individuals are better able to control their subsequent performance than, for instance, a minimum-error computer algorithm. Part of our interest in this question stems from the finding that some individuals perform better on a SAT than on a comparable CAT (Rocklin, 1994; Wise et al. 1992; Wise, Roos, Plake, & Nebelsick-Gullett, 1994). Wise (1994, p. 18), for example, stated "when examinees are allowed to choose their test item difficulty levels, they perceive a sense of control over the test, which serves to reduce anxiety" and which presumably improves performance. Echoing Wise's (1994) thoughts on perceived control, Rocklin et al. (1995, p. 114) stated that "the effects of self-adapted testing can be attributed specifically to the control that examinees exert over the difficulty levels of items they attempt." One explanation of the difference between SAT and CAT versions of the same test is that many examinees experience less anxiety when taking a SAT (Wise et al., 1994). Another explanation, although not mutually exclusive from the reduced anxiety hypothesis, is that some individuals are better able to control their performance than even the most accurate computer-driven selection algorithms. One way to test this difference is to offer good and poor controllers the opportunity to take similar exams using both SAT and CAT formats. Coupled with on-

line or retrospective verbal reports, a comparison between the two methods may illuminate some of the subtle control processes used during testing.

These six questions present an impressive array of topics that warrant further research. Understanding control processes with more precision is important in and of itself. However, understanding the crucial relationship between control and monitoring is even more important, because it is inconceivable that researchers could claim to understand metacognition without understanding the locus and functions of control and monitoring under a wide variety of circumstances, as well as the relationship between them. Similarly, it is essential to understand what makes a highly metacognitive person so able to self-regulate his or her behavior. Comparing good and poor controllers (and monitors) to existing computer software may provide some illustrative insights that increase our understanding, while posing new research questions.

SUMMARY AND CONCLUSIONS

This chapter explored some of the possibilities of using a computer-based testing format to investigate metacognitive processes. We reviewed recent research on control (i.e., regulatory processes used to guide cognitive activities) and monitoring (i.e., regulatory processes used to evaluate the present success of one's performance) functions of metacognition.

After highlighting some of the basic assumptions of computer-based testing, we described several specific strengths of self-adapted testing (SAT). We argued that SAT alleviates a number of serious methodological problems endemic to traditional tests. These included confounds due to differences in ability, prior knowledge, and item difficulty. A more salient problem was that traditional tests do not allow examinees to exert full control over their test-taking behavior. SAT eliminates this problem, and simultaneously offers researchers the opportunity to gather valuable information unobtrusively.

We next considered some of the direct (e.g., item selection time) and indirect (e.g., control accuracy) measures available when using SAT. These measures can be used to answer a host of questions about metacognitive control, as well as the relationship between control and monitoring processes. In addition, it is possible to compare good and poor monitors, as well as to compare the same examinee under CAT and SAT testing conditions. These comparisons offer a unique opportunity to study many aspects of metacognition in a much more direct, yet unobtrusive manner.

Our main conclusion was that computer-based testing formats offer a number of new methodological avenues for the study of metacognition. We proposed six questions that warrant considerable research over the next decade. Chief among these is the relationship between control and monitoring processes, whether these processes share a common pool of resources, and whether they enjoy a reciprocal exchange of information indicative of a regulatory loop. Although little was said concerning developmental issues, we see little difficulty applying these procedures to younger examinees, provided individuals have some knowledge of the test domain, and researchers have access to a calibrated pool of test items.

Finally, despite the tremendous potential of self-adapted testing as a tool for measuring metacognition, we wish to emphasize its essential compatibility with other measurement techniques. SAT seems amenable to on-line and retrospective verbal reports, as well as to on-line subjective performance judgments similar to those used in most monitoring studies. SAT also provides an opportunity to investigate the criteria examinees use to select test items. Concurrent verbal reports may be highly valuable in this regard.

REFERENCES

Alexander, J. M., Carr, M., & Schwanenflugel, P.J. (1995). Development of metacognition in gifted children: Directions for future research. *Developmental Review, 15,* 1-37.

Artzt, A.F., & Armour-Thomas, E. (1992). Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. *Cognition and Instruction, 9,* 137-175.

Assessment Systems Corporation. (1994). User's manual for the MicroCAT Testing System, Version 3.5. St. Paul, MN: Author.

Baker, L. (1989). Metacognition, comprehension monitoring, and the adult reader. *Educational Psychology Review, 1,* 3–38.

Baker, L., & Brown, A. L. (1984). Metacognition and the reading process. In D. Pearson (Ed.), *A handbook of reading research* (pp. 353-394). New York: Plenum Press.

Bjorklund, D. F., & Harnishfeger, K. K. (1990). The resources construct in cognitive development: Diverse resources of evidence and a theory of inefficient inhibition. *Developmental Review, 10,* 48-71.

Borkowski, J., & Muthukrishna, N. (1992). Moving metacognition into the classroom: "Working models" and effective strategy instruction. In M. Pressley, K. Harris, & J. Guthrie (Eds.), *Promoting academic competence and literacy in school* (pp. 477-501). New York: Academic Press.

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. Weinert & R. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65-116). Hillsdale, NJ: Lawrence Erlbaum.

Butterfield, E. C., Nelson, T. O., & Peck, V. (1988). Developmental aspects of the feeling of knowing. *Developmental Psychology, 24,* 654-663.

Cavanaugh, J. C., & Perlmutter, M. (1982). Metamemory: A critical examination. *Child Development, 53,* 11-28.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Holt, Rinehart and Winston, Inc.

Cutler, B.L., & Wolfe, R. N. (1989). Self-monitoring and the association between confidence and accuracy. *Journal of Research in Personality, 23,* 410-420.

Delclos, V. R., & Harrington, C. (1991). Effects of strategy monitoring and proactive instruction on children's problem-solving performance. *Journal of Educational Psychology, 83,* 35-42.

Dixon, R. A., Hultsch, D. F., & Hertzog, C. (1988). The metamemory in adulthood (MIA) questionnaire. *Psychopharmacology Bulletin, 24,* 671-688.

Dunning, D., Griffin, D. W., Milojkovic, J. D., & Ross, L. (1990). The overconfidence effect in social prediction. *Journal of Personality and Social Psychology, 58,* 568-581.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Fischoff, B. (1988). Judgment and decision making. In R. Sternberg & E. Smith (Eds.), *The psychology of human thought* (pp. 153-187). Cambridge, England: Cambridge University Press.

Flavell, J. H. (1987). Speculations about the nature and development of metacognition. In F. Weinert & R. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 21-29). Hillsdale, NJ: Erlbaum.

Funder, D. C. (1987). Errors and Mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101,* 75-90.

Garner, R. (1987). *Metacognition and reading comprehension.* Norwood, NJ: Ablex Publishing.

Garner, R. (1994). Metacognition and executive control. In R. B. Ruddell, M. R. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of reading* (4th ed., pp. 715-732). Newark, DE: International Reading Association.

Garner, R., & Alexander, P. A. (1989). Metacognition: Answered and unanswered questions. *Educational Psychologist, 24,* 143-158.

Ghatala, E. S., Levin, J. R., Foorman, B. R., & Pressley, M. (1989). Improving children's regulation of their reading PREP time. *Contemporary Educational Psychology, 14,* 49-66.

Glaser, R., & Chi, M. T. (1988). Overview. In M. Chi, R. Glaser, & M. Farr (Eds.), *The nature of expertise* (pp. xv-xxviii). Hillsdale, NJ: Erlbaum.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 702-718.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition, 15,* 84-93.

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116,* 119-136.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Jacobs, J. E., & Paris, S. G. (1987). Children's metacognition about reading: Issues in definition, measurement, and instruction. *Educational Psychologist, 22,* 255-278.

Jensen, A. R. (1992). Understanding *g* in terms of information processing. *Educational Psychology Review, 4,* 271-308.

Johnson, M. K., Hastroudi, F., & Lindsay, $. (1994). Source monitoring. *Psychological Review, 101,* 687-698.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217-273.

Kluwe, R. H. (1987). Executive decisions and regulation of problem solving behavior. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 31-64). Hillsdale, NJ: Erlbaum.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100,* 609-639.

Koriat, A. (1994). Memory's knowledge of its own knowledge: The accessibility account of the feeling of knowing. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 115-135). Cambridge, MA: MIT Press.

Kruglanski, A. W. (1989). The psychology of being "right": The problem of accuracy in social prediction and cognition. *Psychological Review, 106,* 395-409.

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 464-470.

Liberman, V., & Tversky, A. (1993). On the evaluation of probability judgments: Calibration, resolution, and monotonicity. *Psychological Bulletin, 114,* 162-173.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallibility during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 756-766.

Lundeberg, M. A., Fox, P. W., & Puncochar, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology, 86,* 114-121.

Maki, R. H., Foley, M. J., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 609-616.

Maki, R. H., & Serra, M. (1992). The basis of test predictions for text materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 116-126.

McDaniel, M. A., & Einstein, G. O. (1989). Material appropriate processing. *Educational Psychology Review, 1,* 113-145.

Metcalfe, J. (1993). Novelty monitoring, metacognition, and control in a composite holographic associative recall model: Implications for Korsakoff amnesia. *Psychological Review, 100,* 3-22.

Meyer, D. E., Irwin, D. E., Osman, A. M., & Kounios, J. (1988). The dynamics of cognition and action: Mental processes inferred from speed-accuracy decomposition. *Psychological Review, 95,* 183-237.

Morris, C. C. (1990). Retrieval processes underlying confidence in comprehension judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 223-232.

Nelson, T. O. (1984). A comparison of current measures of accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109-133.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 450-472). New York: Academic Press.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1-25). Cambridge, MA: MIT Press.

Newman, R. S. (1984). Children's numerical skill and judgments of confidence in estimation. *Journal of Experimental Child Psychology, 37,* 107-123.

Paris, S. G., & Byrnes, J. P. (1989). The constructivist approach to self-regulation and learning in the classroom. In B. Zimmerman & D.

Schunk (Eds.), *Self-regulated learning and academic achievement: Theory, research, and practice* (pp. 169-200). New York: Springer-Verlag.

Pressley, M. (1995). *Advanced educational psychology*. New York: Harper-Collins.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Erlbaum.

Pressley, M., & Ghatala, E. S. (1988). Delusions about performance on multiple-choice comprehension tests. *Reading Research Quarterly, 23,* 454-464.

Pressley, M., & Ghatala, E. S. (1990). Self-regulated learning: Monitoring learning from text. *Educational Psychologist, 25,* 19-33.

Pressley, M., Harris, K. R., & Marks, M. B. (1992). But good strategy instructors are constructivists! *Educational Psychology Review, 4,* 3-31.

Pressley, M., Borkowski, J. G., & Schneider, W. (1987). Cognitive strategies: Good strategies users coordinate metacognition and knowledge. In R. Vasta & G. Whitehurst (Eds.), *Annals of Child Development* (Vol. 5, pp. 89-129). Greenwich, CT: JAI Press.

Pressley, M., Ghatala, E. S., Woloshyn, V., & Pirie, J. (1990). Sometimes adults miss the bid ideas and do not realize it: Confidence in responses to short-answer and multiple-choice comprehension questions. *Reading Research Quarterly, 25,* 232-249.

Pressley, M., Snyder, B. L., Levin, J. R., Murray, H. G., & Ghatala, E. S. (1987). Perceived readiness for examination performance (PREP) produced by initial reading of text and text containing adjunct questions. *Reading Research Quarterly, 22,* 219-236.

Rocklin, T. (1994). Self-adapted testing. *Applied Measurement in Education, 7,* 3-14.

Rocklin, T., & O'Donnell, A. M. (1987). Self-adapted testing: A performance-improving variant of computerized adaptive testing. *Journal of Educational Psychology, 79,* 315-319.

Rocklin, T., & O'Donnell, A. M. (1991, April). *An empirical comparison of self adapted and maximum information item selection.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Rocklin, T. R., O'Donnell, A. M., & Holst, P. M. (1995). Effects and underlying mechanisms of self-adapted testing. *Journal of Educational Psychology, 87,* 103-116.

Roos, L. L., Plake, B. S., & Wise, S. L. (1992, April). *The effects of feedback in computerized adaptive and self-adapted tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Roos, L. L., Wise, S. L., Yoes, M. E., Rocklin, T. R. (in press). Conducting Self-Adapted Testing using MicroCAT. *Educational and Psychological Measurement.*

Schneider, W., & Pressley, M. (1989). *Memory development between 2 and 20.* New York: Springer-Verlag.

Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psychology, 19,* 143-154.

Schraw, G. (1995). *Metacognitve control judgments.* Unpublished data.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19,* 460-475.

Schraw, G., Dunkle, M. E., Bendixen, L., & Roedel, T. D. (1995). Does a general monitoring skill exist? *Journal of Educational Psychology, 87,* 433-444.

Schraw, G., Potenza, M., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology, 18,* 455-463.

Schraw, G., & Roedel, T. D. (1994). Test difficulty and judgment bias. *Memory & Cognition, 22,* 63-69.

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7,* 351-371.

Schwartz, B. L., & Metcalfe, J. (1994). Methodological problems and pitfalls in the study of human metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 93-113). Cambridge, MA: MIT Press.

Slife, B. D., & Weaver, C. A., III. (1992). Depression, cognitive skill, and metacognitive skill in problem solving. *Cognition and Emotion, 6,* 1-22.

Spielberger, C. D. (1980). Preliminary professional manual for the Test Anxiety Inventory. Palo Alto, CA: Consulting Psychologists Press.

Stock, W. A., Kulhavy, R. W., Pridemore, D. R., & Krug, D. (1992). Responding to feedback and multiple choice answers: The influence of response confidence. *Quarterly Journal of Experimental Psychology, 45A,* 649-667.

Swanson, H. L. (1990). Influence of metacognitive knowledge and aptitude on problem solving. *Journal of Educational Psychology, 82,* 306-314.

Thomas, D. L., & Diener, E. (1990). Recalling emotions. *Journal of Personality and Social Psychology, 59,* 291-297.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 4,* 207-232.

Tobias, S. (1996). Interest and metacognitive word knowledge. *Journal of Educational Psychology, 87,* 399-405.

Vispoel, W. P., & Coffman, D. D. (1994). Computer-adaptive and self-adaptive music listening tests: Psychometric features and motivational benefits. *Applied Measurement in Education, 7,* 25-51.

Vispoel, W. P., Rocklin, T., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computer-adaptive, and self-adaptive testing. *Applied Measurement in Education, 7,* 53-79.

Vispoel, W. P., Wang, T., de la Torre, R., Bleiler, T., & Dings, J. (1992, April). *How review options and administration modes influence scores on computerized vocabulary tests.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Weaver, C. A., III. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 214-222.

Weinert, F. E., & Kluwe, R. H. (1987). *Metacognition, motivation, and understanding.* Hillsdale, NJ: Erlbaum.

Willoughby, T., Wood, E., & Khan, M. (1994). Isolating variables that impact or detract from the effectiveness of elaboration strategies. *Journal of Educational Psychology, 86,* 279-289.

Wise, S. L. (1994). Understanding self-adapted testing: The perceived control hypothesis. *Applied Measurement in Education, 7,* 15-24.

Wise, S. L., Plake, B. S., Johnson, P. L., & Roos, L. L. (1992). A comparison of self-adapted and computerized adaptive tests. *Journal of Educational Measurement, 29,* 329-339.

Wise, S. L., Roos, L. L., Plake, B. S., & Nebelsick-Gullet, L. J. (1994). The relationship between examinee anxiety and preference for self-adapted testing. *Applied Measurement in Education, 7,* 81-91.

Wood, E., Pressley, M., & Winne, P. (1990). Elaborative interrogation effects on children's learning of factual content. *Journal of Educational Psychology, 82,* 741-748.

Yan, W. (1994). Learning ability and memory monitoring. *Intelligence, 18,* 215-229.

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice-Hall.